Universidade do Porto

Faculdade de Engenharia

FEUP

Aplicação de ID3 ou C4.5 ao diagnóstico de doença renal crónica

Relatório Intercalar

Inteligência Artificial

3º ano do Mestrado Integrado em Engenharia Informática e Computação

Elementos do grupo:

Luis Oliveira 201304515

Miguel Pereira 201305998

Marta Lopes 201208067

14 de Abril de 2016

Conteúdo

1 Objetivo

Este projeto deverá ser capaz de determinar a árvore de decisão que vai traduzir as regras no diagnóstico da doença renal crónica. Este diagnóstico é feito tendo como base um conjunto de *data sets* a partir dos quais é possivel derivar regras. A árvore de decisão vai ser construída após a derivação e aprendizagem das regras. Finalmente, a árvore poderá ser usada para classificar indivíduos no domínio em análise.

2 Descrição

2.1 Algoritmo C4.5

Decidimos optar por usar este algoritmo em vez do ID3 devido ao facto do algoritmo ID3 apenas aceitar dados discretos, e o C4.5 aceitar dados discretos e contínuos e o nosso *data set* contém dados discretos e contínuos. Este algoritmo é então utilizado para gerar árvores de decisão a partir de um conjunto de dados de treino de amostras já classificadas. Em cada nó da árvore, o algoritmo vai escolher o atributo dos dados que mais efetivamente dividide o seu conjunto de amostras em subconjuntos que tendem para uma categoria ou para outra. O critério de divisão é o ganho de informação normalizado. O atributo com maior ganho é escolhido para tomar a decisão. O algoritmo C4.5 repete assim a etapa anterior nas partições menores.

Este algoritmo vai ter alguns casos base:

- Todas as amostras do conjunto pertencem à mesma categoria: quando este caso ocorre, o algoritmo vai criar um nó folha para a árvore de decisão e escolhe a categoria em questão;
- Nenhuma das características fornece ganho de informação: neste caso, o algoritmo cria um nó de decisão árvore acima usando o valor esperado;
- Instâncias previamente não vistas: novamente, o C4.5 vai criar um nó de decisão árvore acima usando o valor esperado.

2.2 Especificação

Sabendo que iriamos aplicar o algoritmo C4.5 optamos por usar a *framework* See5/C5.0 para realizar o nosso projeto. O C5.0 é uma versão melhorada do C4.5. O algoritmo vai avaliar cada atributo e a partir da avaliação retirada vai criar a árvore de decisão final. Para isso vamos usar um *dataset*

com possíveis portadores e não portadores da doença, para treinar o algoritmo para que aprenda a classificar os novos portadores, usando a árvore de decisão gerada pelo algoritmo.

Em termos de precisão, o C5.0, vai ter um *rate* de erros menor e apesar de que os *rulesets* do C4.5 terem a mesma precisão em termos de previsões, a *ruleset* do C5.0 é menor, o que é vantajoso. O C5.0 vai ser também muito mais rápido e optimizado e a memória que irá ser usada para construir os *rulesets* vai ser muito menor.

O *data set* usado possui um *test data set* e um *training data set*. Este vai conter 280 entradas, das quais 170 são indivíduos diagnosticados com a doença renal crónica e 110 são individuos saudáveis.

2.3 Trabalho efetuado

Inicialmente, obtivemos o *data set* do *site* disponibilizado no enunciado do nosso projeto. O *data set* continha a informação necessária sobre os pacientes para saber se eles seriam diagnosticados com a patologia, ou não. A informação estava toda agregada, por isso foi necessário separar em duas partes: 70% da informação foi para o *training data set* e a outra parte da informação foi para o *test data set*. Para o algoritmo poder reconhecer as colunas e o tipo de dados recolhidos nas amostras, o grupo teve de elaborar o ficheiro *.names* que contém a descrição das mesmas. De seguida, elaboramos o ficheiro *Main.cpp* onde implementamos as funções necessárias para carregar os dados de treino e os dados de teste para a *framework* poder interpretar e devolver os resultados finais de acordo com as *flags* que formos utilizando.

Tendo já todos os dados finais obtidos, vai ser necessário analisá-los e formar as conclusões necessárias para a realização do relatório final.

2.4 Resultados esperados e forma de avaliação

Na fase final deste projeto vai ser necessário implementar uma medição dos resultados obtidos, permitindo assim validar os resultados.

Uma das formas de obter esta validação seria escolher apenas alguns dos indivíduos no *training data set* para a construção da árvore de decisão. Esta árvore deverá ser capaz de classificar corretamente todos os casos de treino, incluindo os indvíduos que não foram usados para a construção da árvore e os indivíduos do *test data set*.

Outra forma de obter a validação consiste em submeter o *training data set* na totalidade no algoritmo de construção da árvore de decisão. Cada entrada será então submetida nessa árvore de decisão e, se tudo estiver bem

implementado, a árvore vai diagnosticar todos os indivíduos do *test data set* com a Doença Renal Crónica.

3 Conclusão

Após a primeira fase de testes ao algoritmo e com a utilização da *framework* See5/C5.0, podemos concluir que é importante proceder a uma análise cuidada e pormenorizada dos atributos e verificar se todos serão necessários para a criação da árvore, para assim levar a uma maior taxa de acertos. É importante poder utilizar a *framework* sem dificuldades e para isso foram criadas as funções necessárias para um fácil carregamento de dados e processamento dos mesmos.

Contudo, podemos afirmar que todo o trabalho desenvolvido até agora contribuiu para o desenvolvimento dos nossos conhecimentos acerca de sistemas de aprendizagem simbólicos automáticos, mais especificamente a aplicação e funcionamento do algoritmo C4.5. Concluimos que este algoritmo tem bastantes potencialidades na ajuda da previsão e deteção da Doença Renal Crónica, e até talvez de outras doenças, de uma forma rápida e eficiente.

4 Recursos

TeXworks: *https://www.tug.org/texworks/*
Github: *https://www.github.com*
See5/C5.0: *https://www.rulequest.com/r210.html*
C5.0: An Informal Tutorial: *https://www.rulequest.com/see5-unix.html*
Wikipédia - Algoritmo C4.5: *https://pt.wikipedia.org/wiki/Algoritmo_C4.5*